M-064E
C.2.

Statistics Canada Statistique Canada

Ottawa

RELIABILITY OF SAMPLING for the MEDICAL CARE PATIENT ORIENTED DATA BASE

K. McClean
Institutional & Agriculture
Survey Methods Division
December, 1981

TABLE OF CONTENTS

INTRODUCTION 1

BACKGROUND 1

- Method One: Coefficient of Variation 2
- Method Two: Minimum Number Present in Sample 12

SUMMARY 16

TABLE I : Insured Population (on June 30) Saskatchewan 1973 .. 1

TABLE IIA : Number of Male Patients, Saskatchewan 1973 3

TABLE IIB : Number of Female Patients, Saskatchewan 1973 4

TABLE IIIA: Percentage of Male Patients, Saskatchewan 1973 5

TABLE IIIB: Percentage of Female Patients, Saskatchewan 1973 ... 6

TABLE IV : Approximate NP Required to Yield Specified C.V. for given P and N 7

TABLE V : Census Population (to nearest thousand) Ontario, 1976 8

TABLE VI : Approximate P (%) Required to Yield Specified C.V. 9

TABLE VII : Sample Size from Infinite Population for 95% Confidence 10

TABLE VIII: Approximate Sampling Rate (%) for Precision of Estimates for Different Population Sizes 11

TABLE IX : (Smallest) Population Sizes (000's) for Different Sampling Rates and Estimates at Chosen Precision Level 12

TABLE X : Approximate NP's Required for Specified m, f, and a's where $Pr(r<m) < \varepsilon$ i.e. $Pr(\frac{r}{n}<\frac{m}{n}) < \varepsilon$

and $Pr(x < a\,\sigma_{\frac{r}{n}}) = \varepsilon$ 14

TABLE XI : P(%) Required for Specified m when f = 0.2 and a = 1.64 16

INTRODUCTION

The Medical Care Patient Oriented Data Base has been proposed to study the services, for different diseases, that an individual receives during his or her lifetime. An individual generates, on average, ten claims a year to the medical insurance commission of which he is a beneficiary. If the proposed data base were to be set up covering every individual insured in Canada, some two hundred and fifty million claims records a year would need to be handled. To reduce this number, it was suggested that samples of beneficiaries be taken in the ten provinces.

The population sizes of the ten provinces vary from 125,000 in Prince Edward Island to 8½ million in Ontario. Five of the ten provinces have populations of less than one million people. The problem of these smaller population provinces is that low frequency diseases and procedures may not be reliably estimated if sampling is undertaken.

This report looks at the problem of estimating low frequency diseases and procedures when sampling is done in the smaller population provinces by considering two different methods. The first method looks at the resulting coefficient of variation when sampling is undertaken while the second method looks at the rate of occurrence of conditions and procedures that would give a specified minimum number in the sample. These two methods described above are discussed separately in sections following.

BACKGROUND

During the years 1972-75, a pilot project was conducted by Statistics Canada with the cooperation of Saskatchewan. Files of claims records received by the Saskatchewan Medical Care Insurance Commission were forwarded to Statistics Canada. Tabulations for 1973 were readily available showing the number of beneficiaries who received at least one service for each ICDA-8 three digit classification and at least one service for each fee schedule classification. Because of this availability of data and the fact that Saskatchewan is one of the smaller population provinces, these 1973 tabulations were used for this report.

Table 1 shows the insured population distribution of Saskatchewan by age and sex on June 30, 1973. All rates for Saskatchewan presented in this report are based on this stable population estimate. The problem of handling continuously changing population sizes is not considered in this report.

TABLE I: Insured Population (on June 30) Saskatchewan 1973

	AGE																	
	Total	Under 1	1-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75&Over
MALES	447,967	7,687	31,964	46,334	50,184	48,234	33,930	25,762	21,817	21,930	23,426	23,893	23,349	22,216	19,116	16,309	11,807	20,009
FEMALES	442,450	7,380	30,313	44,476	48,254	46,983	34,188	26,285	22,481	22,005	23,120	24,131	23,836	21,966	18,160	15,662	11,870	21,340
TOTAL	890,417	15,067	62,277	90,810	98,438	95,217	68,118	52,047	44,298	43,935	46,546	48,024	47,185	44,182	37,276	31,971	23,677	41,349

Some diseases and procedures that are potential areas for studies from the data base were mentioned in meetings of the feasibility project team. Tables IIA and IIB show the number of males and females respectively who received at least one service for these diseases and procedures in 1973. Tables IIIA and IIIB show the calculated percentage of males and females respectively receiving at least one service, that is the number given in Table II divided by the population count given in Table I.

The conditions and procedures presented in these tables have a wide range of frequencies and therefore can represent any condition or procedure that could conceivably be studied from the data base. Complete Assessment by a GP, where the frequencies by age and sex range from fifteen to fifty percent, could represent any highly frequently occuring condition or procedure. At the other end of this scale of occurrence, Alcoholic Psychoses, Multiple Sclerosis, and Appendectomies could represent those conditions and procedures that occur at a frequency less than one half of one percent. The conditions and procedures presented could also represent those which occur equally often through all ages or occur more frequently in the early, middle, or late years of life. Thus while only sixteen conditions and fourteen procedures are presented, a larger spectrum of diseases and procedures are represented.

- Method One: Coefficient of Variation

In simple random sampling, the proportion P of the population or of a particular age/sex subgroup of the population, N, that has a condition of interest is estimated by the poportion p of the sample n that has the condition. The estimated number of individuals in the population having this condition is then $\hat{A} = Np$ with $\mathrm{Var}(\hat{A}) = \frac{N^2P(1-P)}{n}\left(\frac{N-n}{N-1}\right)$. Since the populations that will be studied are quite large, the approximation $N-1 \approx N$ is used to obtain the simpler formula

$$\mathrm{Var}(\hat{A}) = \frac{NP(1-P)(N-n)}{n}$$

$$= \frac{N^2P(1-P)(1-f)}{n}$$

where $f = \frac{n}{N}$ is the sampling rate. The coefficient of variation of $\hat{A}$, which is denoted by C.V., is given by

$$\text{C.V.} = \frac{\sqrt{\mathrm{Var}(\hat{A})}}{NP}$$

$$= \sqrt{\frac{(1-P)(1-f)}{Pn}} \qquad (1)$$

TABLE IIA: Number of Male Patients, Saskatchewan 1973

	CONDITIONS AND PROCEDURES		AGE																
	Code	Description	Total	<1	1-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	≥75
ICDA-8th REVISION	250	Diabetes Millitus	7,193	3	38	75	91	155	155	143	155	226	351	421	580	733	855	1,029	790	1,393
	400-404	Hypertensive Disease	13,348	0	6	15	18	83	154	205	287	350	602	801	1,170	1,506	1,793	1,932	1,652	2,774
	410-414	Ischemic Heart Disease	16,444	0	17	7	8	17	29	64	106	191	463	808	1,386	1,948	2,278	2,463	2,156	4,503
	052	Chickenpox	853	97	298	333	65	19	10	6	9	2	2	1	4	1	3	1	1	1
	056	Rubella	443	186	159	46	34	6	3	4	0	1	0	2	1	0	1	0	0	0
	075	Infectious Mononucleosis	854	6	52	129	129	262	139	54	27	20	14	8	6	2	1	1	2	2
	291	Alcoholic Psychoses	161	0	0	0	0	1	5	7	17	7	25	29	22	17	11	7	5	8
	340	Multiple Sclerosis	290	0	0	2	1	7	6	15	20	27	34	38	44	33	29	19	2	13
	460	Acute Nasopharyngitis (Cold)	5,392	752	1,027	586	387	372	296	226	167	176	176	186	184	192	192	178	113	182
	463	Acute Tonsillitis	12,074	556	3,791	2,644	1,427	1,310	834	424	302	218	164	125	89	53	52	44	23	18
	540-543	Appendicitis	2,303	0	46	373	573	415	216	132	100	91	77	51	53	37	44	42	17	36
	571	Chronic Liver Disease and Cirrhosis	403	1	7	3	5	10	10	4	18	22	55	63	64	69	57	37	25	23
	640-645	Abortions	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	640	Therapeutic Abortion	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	643	Spontaneous Abortion	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	749	Cleft Palate & Cleft Lip	121	22	25	22	31	14	2	2	1	0	1	0	1	0	0	0	0	0
FEE SCHEDULE	3B	Complete Assessment - GP	100,926	2,620	7,249	7,967	8,243	8,593	7,677	5,849	4,833	4,696	5,202	5,567	5,719	5,793	5,495	5,092	3,946	6,385
	87L	Radical Mastectomy	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	241L	Hiatus Hernia Repair	82	0	0	0	0	1	2	1	3	6	3	10	10	12	10	8	10	6
	333L	Appendectomy	1,060	1	16	162	257	199	111	66	45	42	34	22	26	17	23	18	6	15
	499L	Excision or Fulguration of Single Plantar Wart	1,619	0	21	187	357	330	165	87	65	65	70	71	41	29	42	27	21	41
	359M	Arthroscopy	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	320N-329N	Cleft Lip & Palate Services	43	13	19	3	4	1	1	2	0	0	0	0	0	0	0	0	0	0
	35P	Amniotic Tap - Trans-Abdominal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	118P	Dilatation & Curettage	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	120P-125P	Hysterectomies	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	65R-66R	Circumcision-Newborn & Child	4,423	4,175	174	65	9	0	0	0	0	0	0	0	0	0	0	0	0	0
	190R	Vasectomy	2,049	0	0	0	0	0	62	353	554	449	317	184	70	22	9	9	5	15
	145T-146T	Tonsillectomy	2,519	2	772	1,010	293	153	120	78	33	19	10	14	8	3	3	1	0	0
	131A	Submission of Papanicolau Smear	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

TABLE IIB: Number of Female Patients, Saskatchewan 1973

CONDITIONS AND PROCEDURES			AGE																
	Code	Description	Total	<1	1-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	≥75
ICDA-8th REVISION	250	Diabetes Millitus	7,527	5	28	56	98	177	192	174	186	211	298	438	509	701	825	1,015	973	1,641
	400-404	Hypertensive Disease	25,952	0	8	9	29	95	155	210	264	418	834	1,500	2,294	3,046	3,445	4,024	3,583	6,038
	410-414	Ischemic Heart Disease	11,696	0	8	6	14	17	27	52	43	100	203	394	630	960	1,368	1,755	1,844	4,275
	052	Chickenpox	802	119	251	296	65	23	17	5	12	5	2	1	1	1	2	0	1	1
	056	Rubella	519	156	167	54	36	24	32	26	13	3	4	0	3	0	0	0	1	0
	075	Infectious Mononucleosis	958	0	32	109	168	351	136	68	33	15	9	11	11	6	4	3	1	1
	291	Alcoholic Psychoses	44	0	0	0	0	3	3	1	4	7	6	4	4	7	2	0	2	1
	340	Multiple Sclerosis	420	0	0	1	2	18	28	41	34	36	56	47	62	40	25	14	7	9
	460	Acute Nasopharyngitis (Cold)	6,192	700	965	575	452	614	500	370	256	245	222	230	229	226	174	159	105	170
	463	Acute Tonsillitis	13,340	409	3,178	2,963	2,028	1,986	906	573	331	220	172	154	125	107	64	69	25	30
	540-543	Appendicitis	2,500	1	39	339	606	598	306	164	130	65	59	53	38	33	21	18	11	19
	571	Chronic Liver Disease and Cirrhosis	202	2	2	3	4	11	5	9	10	11	22	23	15	24	22	11	13	15
	640-645	Abortions	2,518	0	0	0	32	573	768	581	282	161	100	19	2	0	0	0	0	0
	640	Therapeutic Abortion	788	0	0	0	21	292	226	110	54	51	27	7	0	0	0	0	0	0
	643	Spontaneous Abortion	186	0	0	0	1	28	62	62	20	10	3	0	0	0	0	0	0	0
	749	Cleft Palate & Cleft Lip	83	23	21	12	20	3	2	0	0	0	0	0	0	1	0	0	1	0
FEE SCHEDULE	3B	Complete Assessment - GP	139,929	2,324	6,802	7,465	7,922	15,633	15,058	10,736	8,758	8,190	8,582	9,037	8,757	7,797	6,419	5,531	4,223	6,695
	87L	Radical Mastectomy	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
	241L	Hiatus Hernia Repair	84	0	0	0	0	0	0	0	4	3	3	8	8	14	24	11	6	3
	333L	Appendectomy	1,042	0	16	132	257	260	129	71	55	20	29	21	16	13	7	9	1	6
	499L	Excision or Fulguration of Single Plantar Wart	2,142	0	26	215	482	495	243	134	91	82	81	80	54	47	38	30	21	23
	359M	Arthroscopy	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
	320N-329N	Cleft Lip & Palate Services	37	16	12	4	3	2	0	0	0	0	0	0	0	0	0	0	0	0
	35P	Amniotic Tap - Trans-Abdominal	111	0	0	0	0	16	24	33	18	13	7	0	0	0	0	0	0	0
	118P	Dilatation & Curettage	7,082	0	0	0	21	473	1,052	1,088	855	753	773	863	622	264	138	77	65	38
	120P-125P	Hysterectomies	1,740	0	0	0	2	9	26	85	185	324	384	377	211	50	38	22	18	9
	65R-66R	Circumcision-Newborn & Child	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	190R	Vasectomy	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	145T-146T	Tonsillectomy	2,895	0	585	1,100	493	387	172	69	28	19	13	6	11	6	3	2	1	0
	131A	Submission of Papanicolau Smear	93,351	8	15	21	152	7,184	15,335	12,940	10,260	8,926	8,492	8,293	7,035	5,423	3,905	2,595	1,495	1,272

TABLE IIIA: Percentage of Male Patients, Saskatchewan 1973

	CONDITIONS AND PROCEDURES		AGE																
	Code	Description	Total	<1	1-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	≥75
			%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
ICDA-8th REVISION	250	Diabetes Millitus	1.60	0.04	0.12	0.16	0.18	0.32	0.46	0.56	0.71	1.03	1.50	1.76	2.48	3.30	4.47	6.31	6.69	6.96
	400-404	Hypertensive Disease	2.98	0	0.02	0.03	0.04	0.17	0.45	0.80	1.32	1.60	2.57	3.35	5.01	6.78	9.38	11.85	13.99	13.86
	410-414	Ischemic Heart Disease	3.67	0	0.05	0.02	0.02	0.04	0.09	0.25	0.49	0.87	1.98	3.38	5.94	8.77	11.92	15.10	18.26	22.50
	052	Chickenpox	0.19	1.26	0.93	0.72	0.13	0.04	0.03	0.02	0.04	0.01	0.01	0	0.02	0	0.02	0.01	0.01	0
	056	Rubella	0.10	2.42	0.50	0.10	0.07	0.01	0.01	0.02	0	0	0	0.01	0	0	0.01	0	0	0
	075	Infectious Mononucleosis	0.19	0.08	0.16	0.28	0.26	0.54	0.41	0.21	0.12	0.09	0.06	0.03	0.03	0.01	0.01	0.01	0.02	0.01
	291	Alcoholic Psychoses	0.04	0	0	0	0	0	0.01	0.03	0.08	0.03	0.11	0.12	0.09	0.08	0.06	0.04	0.04	0.04
	340	Multiple Sclerosis	0.06	0	0	0	0	0.01	0.02	0.06	0.09	0.12	0.15	0.16	0.19	0.15	0.15	0.12	0.02	0.06
	460	Acute Nasopharyngitis (Cold)	1.20	9.78	3.21	1.26	0.77	0.77	0.87	0.88	0.77	0.80	0.75	0.78	0.79	0.86	1.00	1.09	0.96	0.91
	463	Acute Tonsillitis	2.70	7.23	11.86	5.71	2.84	2.72	2.46	1.65	1.38	0.99	0.70	0.52	0.38	0.24	0.27	0.27	0.19	0.09
	540-543	Appendicitis	0.51	0	0.14	0.81	1.14	0.86	0.64	0.51	0.46	0.41	0.33	0.21	0.23	0.17	0.23	0.26	0.14	0.18
	571	Chronic Liver Disease and Cirrhosis	0.09	0.01	0.02	0.01	0.01	0.02	0.03	0.02	0.08	0.10	0.23	0.26	0.27	0.31	0.30	0.23	0.21	0.11
	640-645	Abortions	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	640	Therapeutic Abortion	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	643	Spontaneous Abortion	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	749	Cleft Palate & Cleft Lip	0.03	0.29	0.08	0.05	0.06	0.03	0.01	0.01	0	0	0	0	0	0	0	0	0	0
FEE SCHEDULE	3B	Complete Assessment - GP	22.53	34.08	22.68	17.19	16.43	17.82	22.63	22.70	22.15	21.41	22.21	23.30	24.49	26.08	28.75	31.22	33.42	31.91
	87L	Radical Mastectomy	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	241L	Hiatus Hernia Repair	0.02	0	0	0	0	0	0.01	0	0.01	0.03	0.01	0.04	0.04	0.05	0.05	0.05	0.08	0.03
	333L	Appendectomy	0.24	0.01	0.05	0.35	0.51	0.41	0.33	0.26	0.21	0.19	0.15	0.09	0.11	0.08	0.12	0.11	0.05	0.07
	499L	Excision or Fulguration of Single Plantar Wart	0.36	0	0.07	0.40	0.71	0.68	0.49	0.34	0.30	0.30	0.30	0.30	0.18	0.13	0.22	0.17	0.18	0.20
	359M	Arthroscopy	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	320N-329N	Cleft Lip & Palate Services	0.01	0.17	0.06	0.01	0.01	0	0	0.01	0	0	0	0	0	0	0	0	0	0
	35P	Amniotic Tap - Trans-Abdominal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	118P	Dilatation & Curettage	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	120P-125P	Hysterectomies	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	65R-66R	Circumcision-Newborn & Child	0.99	54.31	0.54	0.14	0.02	0	0	0	0	0	0	0	0	0	0	0	0	0
	190R	Vasectomy	0.46	0	0	0	0	0	0.18	1.37	2.54	2.05	1.35	0.77	0.30	0.10	0.05	0.06	0.04	0.07
	145T-146T	Tonsillectomy	0.56	0.03	2.42	2.18	0.58	0.32	0.35	0.30	0.15	0.09	0.04	0.06	0.03	0.01	0.02	0.01	0	0
	131A	Submission of Papanicolau Smear	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

TABLE IIIB: Percentage of Female Patients, Saskatchewan 1973

CONDITIONS AND PROCEDURES			AGE																
	Code	Description	Total	<1	1-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75
			%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%	%
ICDA-8th REVISION	250	Diabetes Millitus	1.70	0.07	0.09	0.13	0.20	0.38	0.56	0.66	0.83	0.96	1.29	1.82	2.14	3.19	4.54	6.48	8.20	7.69
	400-404	Hypertensive Disease	5.87	0	0.03	0.02	0.06	0.20	0.45	0.80	1.17	1.90	3.61	6.22	9.62	13.87	18.97	25.69	30.19	28.29
	410-414	Ischemic Heart Disease	2.64	0	0.03	0.01	0.03	0.04	0.08	0.20	0.19	0.45	0.88	1.63	2.64	4.37	7.53	11.21	15.53	20.03
	052	Chickenpox	0.18	1.61	0.83	0.67	0.13	0.05	0.05	0.02	0.05	0.02	0.01	0	0	0	0.01	0	0.01	0
	056	Rubella	0.12	2.11	0.55	0.12	0.07	0.05	0.09	0.10	0.06	0.01	0.02	0	0.01	0	0	0	0.01	0
	075	Infectious Mononucleosis	0.22	0	0.11	0.25	0.35	0.75	0.40	0.26	0.15	0.07	0.04	0.05	0.05	0.03	0.02	0.02	0.01	0
	291	Alcoholic Pyschoses	0.01	0	0	0	0	0.01	0.01	0	0.02	0.03	0.03	0.02	0.02	0.03	0.01	0	0.02	0
	340	Multiple Sclerosis	0.09	0	0	0	0	0.04	0.08	0.16	0.15	0.16	0.24	0.19	0.26	0.18	0.14	0.09	0.06	0.04
	460	Acute Nasopharyngitis (Cold)	1.40	9.49	3.18	1.29	0.94	1.31	1.46	1.41	1.14	1.11	0.96	0.95	0.96	1.03	0.96	1.02	0.88	0.80
	463	Acute Tonsillitis	3.02	5.54	10.48	6.66	4.20	4.23	2.65	2.18	1.47	1.00	0.74	0.64	0.52	0.49	0.35	0.44	0.21	0.14
	540-543	Appendicitis	0.57	0.01	0.13	0.76	1.26	1.27	0.90	0.62	0.58	0.30	0.26	0.22	0.16	0.15	0.12	0.11	0.09	0.09
	571	Chronic Liver Disease and Cirrhosis	0.05	0.03	0.01	0.01	0.01	0.02	0.01	0.03	0.04	0.05	0.10	0.10	0.06	0.11	0.12	0.07	0.11	0.07
	640-645	Abortions	0.57	0	0	0	0.07	1.22	2.25	2.21	1.25	0.73	0.43	0.08	0.01	0	0	0	0	0
	640	Therapeutic Abortion	0.18	0	0	0	0.04	0.62	0.66	0.42	0.24	0.23	0.12	0.03	0	0	0	0	0	0
	643	Spontaneous Abortion	0.04	0	0	0	0	0.06	0.18	0.24	0.09	0.05	0.01	0	0	0	0	0	0	0
	749	Cleft Palate & Cleft Lip	0.02	0.31	0.07	0.03	0.04	0.01	0.01	0	0	0	0	0	0	0	0	0	0.01	0
FEE SCHEDULE	3B	Complete Assessment - GP	31.63	31.49	22.44	16.78	16.42	33.27	44.04	40.84	38.96	37.22	37.12	37.45	36.74	35.50	35.35	35.31	35.58	31.37
	87L	Radical Mastectomy	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0	0	0
	241L	Hiatus Hernia Repair	0.02	0	0	0	0	0	0	0	0.02	0.01	0.01	0.03	0.03	0.06	0.13	0.07	0.05	0.01
	333L	Appendectomy	0.24	0	0.05	0.30	0.53	0.55	0.38	0.27	0.24	0.09	0.13	0.09	0.07	0.06	0.04	0.06	0.01	0.03
	499L	Excision or Fulguration of Single Plantar Wart	0.48	0	0.09	0.48	1.00	1.05	0.71	0.51	0.40	0.37	0.35	0.33	0.23	0.21	0.21	0.19	0.18	0.11
	359M	Arthroscopy	0	0	0	0	0	0.001	0	0	0	0	0	0	0	0	0	0	0	0
	320N-329N	Cleft Lip & Palate Services	0.01	0.22	0.04	0.01	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0
	35P	Amniotic Tap - Trans-Abdominal	0.03	0	0	0	0	0.03	0.07	0.13	0.08	0.06	0.03	0	0	0	0	0	0	0
	118P	Dilatation & Curettage	1.60	0	0	0	0.04	1.01	3.08	4.14	3.80	3.42	3.34	3.58	2.61	1.20	0.76	0.49	0.55	0.18
	120P-125P	Hysterectomies	0.39	0	0	0	0	0.02	0.08	0.32	0.82	1.47	1.66	1.56	0.89	0.23	0.21	0.14	0.15	0.04
	65R-66R	Circumcision-Newborn & Child.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	190R	Vasectomy	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	145T-146T	Tonsillectomy	0.65	0	1.93	2.47	1.02	0.82	0.50	0.26	0.12	0.09	0.06	0.02	0.05	0.03	0.02	0.01	0.01	0
	131A	Submission of Papanicolau Smear	21.10	0.11	0.05	0.05	0.32	15.29	44.85	49.23	45.64	40.56	36.73	34.37	29.51	24.67	21.50	16.57	12.59	5.96

Solving equation (1) for P yields $P = \left[1 + \frac{n(C.V.)^2}{(1-f)}\right]^{-1}$. (2)

The proportion P of the population N having the condition of interest that would yield a specified coefficient of variation, if sampling was done at some rate f, can be calculated from equation (2). Table IV shows the appropriate number of individuals NP in the population N required to guarantee that the resulting coefficient of variation, when sampling at some rate f, would be the one specified. For example, if a population of N = 100,000 was sampled at a rate f = 0.10, to achieve a coefficient of variation of 0.1 or less, the number of individuals having this condition in the total population would have to be at least 890 of the 100,000.

TABLE IV: Approximate NP Required to Yield Specified C.V. for given f and N

N	f = .05			f = .1			f = .2			f = .5		
	CV=.05	CV=.1	CV=.2	CV=.05	CV=.1	CV=.2	CV=.05	CV=.1	CV=.2	CV=.05	CV=.1	CV=.2
5,000	3015	1375	435	2095	765	215	1210	370	100	370	100	25
10,000	4320	1595	455	2645	825	220	1380	385	100	385	100	25
15,000	5045	1685	460	2905	850	220	1445	390	100	390	100	25
20,000	5510	1735	465	3050	860	220	1480	390	100	390	100	25
25,000	5830	1765	465	3145	870	225	1505	395	100	395	100	25
30,000	6065	1785	470	3215	875	225	1520	395	100	395	100	25
35,000	6245	1800	470	3265	875	225	1530	395	100	395	100	25
40,000	6385	1815	470	3305	880	225	1540	395	100	395	100	25
45,000	6500	1825	470	3335	880	225	1545	395	100	395	100	25
50,000	6600	1830	470	3360	885	225	1550	395	100	395	100	25
100,000	7065	1865	475	3475	890	225	1575	400	100	400	100	25
150,000	7235	1875	475	3515	895	225	1585	400	100	400	100	25
200,000	7320	1880	475	3535	895	225	1585	400	100	400	100	25
250,000	7375	1885	475	3550	895	225	1590	400	100	400	100	25
300,000	7410	1890	475	3555	895	225	1590	400	100	400	100	25
350,000	7440	1890	475	3565	900	225	1595	400	100	400	100	25
400,000	7460	1890	475	3570	900	225	1595	400	100	400	100	25
450,000	7475	1890	475	3570	900	225	1595	400	100	400	100	25

The coefficient of variation is a measure of the reliability of the estimate $\hat{A}$. Estimates giving C.V. < 0.05, that is the estimate is within 5% of the true number, are quite reliable whereas estimates giving C.V. > 0.2, that is the estimate is not within 20% of the true number, are fairly unreliable. By comparing Table IV to Tables IIA and IIB, it can be seen which type of condition or procedure will yield reliable estimates if sampling was undertaken in Saskatchewan.

Conditions or procedures that occur as frequently as Complete Assessments will have coefficients of variation less than 0.05 even for sampling rates as small as 5%. Conditions or procedures that occur as infrequently as Multiple Sclerosis will have coefficients of variation greater than 0.2 even for sampling rates as large as 50%.

For a large number of the conditions and procedures that were considered, a sampling rate of 5% would not result in a coefficient of variation less than 0.10 even if the total population of either males or females, without any breakdown into age groups, were to be considered. Since age is a factor of great interest in medical studies, this low rate of sampling is not advisable.

Sampling in Saskatchewan at rates of 10% and 20% is better than at a rate of 5% but there still are conditions and procedures that are so infrequent that a coefficient of variation of 0.2 cannot be achieved even for the total population of either males or females. In a number of the areas of interest presented, grouping ages into 10 or 20 year age groups will achieve C.V. < 0.10 for sampling at 20%, but only the more frequently occuring conditions and procedures will achieve C.V. < 0.10 in those age groups for sampling at 10%.

Sampling in Saskatchewan at a rate of 50% will achieve C.V. < 0.20 for those conditions and procedures that occur as infrequently as Cleft Lip and Cleft Palate Services if the total population of either males or females is considered. For other conditions and procedures, grouping by five to ten year age groups appears to be sufficient to yield C.V. < 0.10.

This method suggests that sampling in the six smaller population provinces at a rate of 50% will not give reliable estimates, for small age groups, for the wide range of conditions and procedures that might be studied. Thus, sampling should not be conducted in these provinces.

Consider the large province of Ontario. Its 1976 population distribution is shown in Table V. Using equation (2), Table VI is produced to show the approximate proportion of the population N required to guarantee the specified coefficient of variation for a few population sizes and sampling rates. If it is assumed that the rate of occurrence of conditions and procedures in Ontario is the same as the rate of occurrence in Saskatchewan, a comparison of Table VI with Table III will indicate that sampling is more reliable for the province of Ontario. For a sampling rate of 20% in Ontario, conditions and procedures that occur as infrequently as Alcoholic Psychoses, Multiple Sclerosis and Appendectomies will have coefficients of variation less than 0.2 for five year age groups and in some, the coefficient of variation will be less than 0.1. This shows that although sampling is not advisable for provinces whose population is smaller than that of Saskatchewan, it will be feasible for the larger population provinces.

TABLE V: Census Population (to nearest thousand) Ontario, 1976

	AGE																	
	Total	Under 1	1-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75&Over
MALES	4,097	61	251	342	409	413	368	356	296	247	241	239	227	179	157	121	86	103
FEMALES	4,168	58	237	326	389	395	376	359	291	243	234	233	238	190	169	140	111	177
TOTAL	8,264	119	488	668	799	808	744	716	586	490	475	472	465	369	326	261	197	281

TABLE VI: Approximate P (%) Required to Yield Specified C.V.

N	f = .1			f = .2		
	CV=.05	CV=.1	CV=.2	CV=.05	CV=.1	CV=.2
100,000	3.47%	0.89%	0.22%	1.57%	0.40%	0.10%
200,000	1.77	0.45	0.11	0.79	0.20	0.05
300,000	1.19	0.30	0.07	0.53	0.13	0.03
400,000	0.89	0.22	0.06	0.40	0.10	0.02

There is another approach, taken from the paper "Adequacy of Sample Size" by the Health Statistical Methodology Unit, World Health Organization, Geneva, that can be used to verify the results obtained using method one. In this approach, the sampling rate required for sampling from an infinite population is found and then a correction factor is applied to this rate to determine the sampling rate required for sampling from the finite population under consideration. The three considerations that must be specified to determine the required sample size are:

1. the precision wanted in the estimate (i.e. the percentage around the estimate to be tolerated)
2. the confidence level wanted (e.g. 95% confidence that the tolerated precision has been attained)

and 3. an apriori guess of what percentage of the population has the characteristic under study

Table VII shows the tolerated range around the true percentage for fifteen values and the sample size required from an infinite population to be 95% confident that the true value is in that range. By applying the correction factor to the numbers in Table VII, the approximate sampling rates required to achieve the desired precision around an estimate for different finite population sizes were found and are presented in Table VIII. For example, to be 95% confident that the true value is within 5% of an estimated rate of 10% for a population of 40,000, the required sampling rate would be 35% and the sample size would be 0.35 x 40,000 = 14,000.

To see the smallest sufficient finite population size necessary to achieve the desired precision about estimates if sampling was done at different sampling rates, Table VIII is rearranged to give Table IX. For example, to be 95% confident that the true value is within 10% of the estimate 0.05, the population size must be at least 25,000 if sampling was to be done at the rate of 25%.

The WHO report did not specify any sample sizes greater than 100,000 and so for these cases a correction factor could not be applied to

TABLE VII: Sample Size from Infinite Population for 95% Confidence

	1% Precision		5% Precision		10% Precision	
Estimates	Range	Infinite Sample Size	Range	Infinite Sample Size	Range	Infinite Sample Size
.001	.00099-.00101	>100,000	.00095-.00105	>100,000	.0009-.0011	>100,000
.005	.00495-.00505	>100,000	.00475-.00525	>100,000	.0045-.0055	90,000
.01	.0099-.0101	>100,000	.0095-.0105	>100,000	.009-.011	40,000
.02	.0198-.0202	>100,000	.019-.021	80,000	.018-.022	20,000
.05	.0495-.0505	>100,000	.0475-.0525	30,000	.045-.055	7,000
.07	.0693-.0707	>100,000	.0665-.0735	25,000	.063-.077	5,000
.1	.099-.101	>100,000	.095-.105	20,000	.09-.11	4,000
.15	.1485-.1515	>100,000	.1425-.1575	8,000	.135-.165	2,000
.2	.198-.202	>100,000	.19-.21	5,000	.18-.22	2,000
.25	.2475-.2525	>100,000	.2375-.2625	5,000	.225-.275	1,000
.3	.297-.303	90,000	.285-.315	4,000	.27-.33	1,000
.4	.396-.404	60,000	.38-.42	3,000	.36-.44	600
.5	.495-.505	40,000	.475-.525	2,000	.45-.55	400
.7	.693-.707	20,000	.665-.735	700	.63-.77	200
.9	.891-.909	5,000	.855-.945	300	.81-.99	80

find the sampling rate for the corresponding finite population. In Table VIII, this fact is indicated in any box showing a sampling rate of >50%. As a result of this, determining what the smallest population size would be for all sampling rates for the estimate 0.001 was impossible, as well as for estimates less than 0.01 with 5% precision and estimates less than 0.25 with 1% precision. All that can be said is that these population sizes would be greater than those for larger estimates.

For the precision level of 10%, it appears that sampling could be considered for estimates of 0.005 or larger. If sampling was done at a rate of 15% or larger, estimates of 0.07 or greater could be estimated to 10% precision at 95% confidence for population sizes of at least 40,000, that is for the five year age groups in Saskatchewan. Even estimates as small as 0.005 could be estimated for populations of size 100,000 if sampling was done at a rate of 50%. In Saskatchewan, this would mean looking at twenty to twenty-five year age groups. However, Table III shows that a number of conditions and procedures occur at rates smaller than 0.5% and thus would not attain the 10% precision level if sampling was done at a 50% rate from populations of 100,000.

This second approach reaches the same conclusions as were reached using method one, that sampling could be done for provinces the size of Saskatchewan if only the relatively high frequency conditions and procedures were to be studied. Those that occur as infrequently as Infectious Mononucleosis, Multiple Sclerosis or Appendectomies could not be studied even if sampling was done at a 50% rate.

TABLE VIII: Approx. Sampling Rate (%) for Precision of Estimates for Different Population Sizes

Estimate	Precision	N (000's)																
		5	10	15	20	25	30	35	40	45	50	100	150	200	250	300	350	400	450
.001	1%	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50							
	5%	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50							
	10%	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	>40	>35	>30	>25	>22½	>20	>20
.005	1%	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50							
	5%	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50							
	10%	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	50	40	32½	27½	25	22½	20	17½
.01	1%	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50							
	5%	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	>40	>35	>30	>25	>22½	>20	>20
	10%	>50	>50	>50	>50	>50	>50	>50	50	47½	45	30	22½	17½	15	12½	12½	10	10
.02	1%	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50							
	5%	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	45	37½	30	25	22½	20	17½	17½
	10%	>50	>50	>50	50	45	40	37½	35	32½	30	17½	12½	10	7½	7½	7½	5	5
.05	1%	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50							
	5%	>50	>50	>50	>50	>50	50	47½	45	40	40	25	17½	15	12½	10	10	7½	7½
	10%	>50	42½	32½	27½	22½	20	17½	15	15	12½	7½	5	5	5	2½	2½	2½	2½
.07	1%	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50							
	5%	>50	>50	>50	>50	50	47½	42½	40	37½	35	20	15	12½	10	10	7½	7½	7½
	10%	50	35	25	20	17½	15	12½	12½	10	10	5	5	2½	2½	2½	2½	2½	2½
.1	1%	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50							
	5%	>50	>50	>50	50	45	40	37½	35	32½	30	17½	12½	10	7½	7½	7½	5	5
	10%	45	30	22½	17½	15	12½	12½	10	10	7½	5	5	2½	2½	2½	2½	1	1
.15	1%	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50							
	5%	>50	45	37½	30	25	22½	20	17½	17½	15	7½	7½	5	5	5	2½	2½	2½
	10%	30	17½	12½	10	7½	7½	7½	5	5	5	2½	2½	1	1	1	1	1	1
.2	1%	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50							
	5%	50	35	25	20	17½	15	12½	12½	10	10	5	5	2½	2½	2½	2½	2½	2½
	10%	30	17½	12½	10	7½	7½	7½	5	5	5	2½	2½	1	1	1	1	1	1
.25	1%	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	>40	>35	>30	>25	>22½	>20	>20
	5%	50	35	25	20	17½	15	12½	12½	10	10	5	5	2½	2½	2½	2½	2½	2½
	10%	17½	10	7½	5	5	5	5	2½	2½	2½	1	1	1	1	1	1	1	1
.3	1%	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	50	40	32½	27½	25	22½	20	17½
	5%	45	30	22½	17½	15	12½	12½	10	10	7½	5	5	2½	2½	2½	2½	1	1
	10%	17½	10	7½	5	5	5	5	2½	2½	2½	1	1	1	1	1	1	1	1
.4	1%	>50	>50	>50	>50	>50	>50	>50	>50	>50	>50	40	30	25	20	17½	15	15	12½
	5%	40	25	17½	15	12½	10	10	7½	7½	7½	5	2½	2½	2½	1	1	1	1
	10%	12½	7½	5	5	2½	2½	2½	2½	2½	2½	1	1	1	1	1	1	1	1
.5	1%	>50	>50	>50	>50	>50	>50	>50	50	47½	45	30	22½	17½	15	12½	12½	10	10
	5%	30	17½	12½	10	7½	7½	7½	5	5	5	2½	2½	1	1	1	1	1	1
	10%	7½	5	5	2½	2½	2½	2½	1	1	1	1	1	1	1	1	1	1	1
.7	1%	>50	>50	>50	50	45	40	37½	35	32½	30	17½	12½	10	7½	7½	7½	5	5
	5%	12½	7½	5	5	5	2½	2½	2½	2½	2½	1	1	1	1	1	1	1	1
	10%	5	2½	2½	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
.9	1%	50	35	25	20	17½	15	12½	12½	10	10	5	5	2½	2½	2½	2½	2½	2½
	5%	7½	5	2½	2½	2½	1	1	1	1	1	1	1	1	1	1	1	1	1
	10%	2½	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

TABLE IX: (Smallest) Population Sizes (000's) for Different Sampling Rates and Estimates at Chosen Precision Level

Precision Level	Precision	Sampling Rate	ESTIMATES														
			.001	.005	.01	.02	.05	.07	.1	.15	.2	.25	.3	.4	.5	.7	.9
	10%	5%				400	150	100	100	40	40	20	20	15	10	5	5
		10%			400	200	100	45	40	20	20	10	10	10	5		
		15%			250	150	40	30	25	15	15			5			
		20%		400	200	100	30	20	20	10	10	5	5				
		25%		300	150		25	15	15								
		30%		250	100	50	20		10	5	5						
		35%		200		40	15	10									
		40%		150		30											
		45%		150	50	25	10		5								
		50%		100	40	20		5									
	5%	5%							400	200	100	100	100	100	40	15	10
		10%					300	250	200	100	45	45	40	30	20	10	5
		15%					200	150	150	50	30	30	25	20	15	5	
		20%				350	150	100	100	35	20	20	20	15	10		
		25%				250	100			25	15	15	15	10			
		30%				200			50	20			10		5		
		35%						50	40		10	10					
		40%				150	45	40	30	15				5			
		45%				100	40	35	25	10			5				
		50%					30	25	20		5	5					
	1%	5%														400	100
		10%													400	200	45
		15%												350	250	150	30
		20%											400	250	200	100	20
		25%											300	200	150		15
		30%											250	150	100	50	
		35%											200			40	10
		40%											150	100		30	
		45%													50	25	
		50%											100		40	20	5

- Method Two: Minimum Number Present in Sample

This method will determine P, the proportion having a condition of interest of the population N, such that when sampling is done at rate f, the probability of having less than m individuals having the condition out of the sample n is some small number. Let R and r be the number having the condition in the total population and sample respectively. For large enough n, it can be shown that $\frac{r}{n} \sim \text{Normal}\left(P, \frac{P(1-P)}{n}\left(\frac{N-n}{N-1}\right)\right)$. Since the populations under consideration are large, the approximation $N-1 \approx N$ can be used to obtain $\frac{r}{n} \sim \text{Normal}\left(P, \frac{P(1-P)(1-f)}{n}\right)$. To satisfy

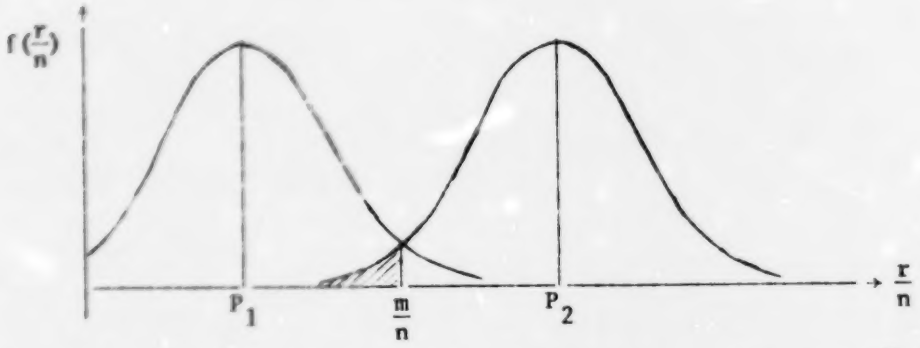
the condition $Pr(r<m) < \varepsilon$ or $Pr(\frac{r}{n} < \frac{m}{n}) < \varepsilon$, find P such that

$$\frac{m}{n} = P - a\sqrt{\frac{P(1-P)(1-f)}{n}} \qquad (3)$$

where $Pr(X<a\,\sigma_{\frac{r}{n}}) = \varepsilon$ and $X \sim$ Normal (0,1). Solving equation (3) for P yields

$$P = \frac{\frac{2m}{n} + \frac{a^2(1-f)}{n} \pm \frac{a}{n}\sqrt{a^2(1-f)^2 + 4m(1-f)(1-\frac{m}{n})}}{2(1 + \frac{a^2(1-f)}{n})} . \qquad (4)$$

This equation will yield two values for P, say P_1 and P_2 with $P_2 > P_1$. Consider the distribution of $\frac{r}{n}$.



The area shaded is the probability discussed above, that is $Pr(\frac{r}{n} < \frac{m}{n}) = \varepsilon$. Hence P_2 is the value of interest and the plus sign is to be used in equation (4).

Consider the following two values for ε, 0.05 and 0.025. These two values determine two values for a as follows:

$$\varepsilon = 0.05 \Rightarrow Pr(X < a\,\sigma_{\frac{r}{n}}) = 0.05 \Rightarrow a = 1.64$$

$$\varepsilon = 0.025 \Rightarrow Pr(X < a\,\sigma_{\frac{r}{n}}) = 0.025 \Rightarrow a = 1.96$$

Using these two values for a and different values for m, f and N, P can be calculated for the different combinations from equation (4). Table X gives the values of NP, the number in the population under study having the condition of interest required to guarantee, to some probability, that at least m individuals in the sample will have this condition of interest.

TABLE X: Approximate NP's Required for Specified m, f, and a's

where $Pr(r<m) < \varepsilon$ i.e. $Pr(\frac{r}{n} < \frac{m}{n}) < \varepsilon$

and $Pr(x < a\,\sigma_{\frac{r}{n}}) = \varepsilon$

a	N	f = .2			f = .5		
		m = 100	m = 500	m=1000	m = 100	m = 500	m=1000
1.64	5,000	575	2,615	5,000	225	1,045	2,060
	10,000	575	2,645	5,165	225	1,050	2,065
	15,000	575	2,655	5,190	225	1,050	2,070
	20,000	580	2,655	5,205	225	1,050	2,070
	25,000	580	2,660	5,210	225	1,050	2,070
	30,000	580	2,660	5,215	225	1,050	2,070
	35,000	580	2,665	5,220	225	1,050	2,070
	40,000	580	2,665	5,220	225	1,050	2,070
	45,000	580	2,665	5,225	225	1,050	2,070
	50,000	580	2,665	5,225	225	1,050	2,070
	100,000	580	2,670	5,230	225	1,055	2,075
	150,000	580	2,670	5,235	225	1,055	2,075
	200,000	580	2,670	5,235	225	1,055	2,075
	250,000	580	2,670	5,235	225	1,055	2,075
	300,000	580	2,670	5,235	225	1,055	2,075
	350,000	580	2,670	5,235	225	1,055	2,075
	400,000	580	2,670	5,235	225	1,055	2,075
	450,000	580	2,670	5,235	225	1,055	2,075
1.96	5,000	590	2,640	5,000	230	1,055	2,070
	10,000	595	2,675	5,190	230	1,060	2,080
	15,000	595	2,685	5,230	230	1,060	2,085
	20,000	595	2,690	5,245	230	1,060	2,085
	25,000	595	2,690	5,255	230	1,060	2,085
	30,000	595	2,695	5,260	230	1,060	2,085
	35,000	595	2,695	5,260	230	1,060	2,085
	40,000	595	2,700	5,265	230	1,060	2,085
	45,000	595	2,700	5,265	230	1,060	2,085
	50,000	595	2,700	5,270	230	1,060	2,085
	100,000	595	2,700	5,280	230	1,065	2,090
	150,000	595	2,700	5,280	230	1,065	2,090
	200,000	595	2,700	5,280	230	1,065	2,090
	250,000	595	2,700	5,280	230	1,065	2,090
	300,000	595	2,700	5,280	230	1,065	2,090
	350,000	595	2,700	5,280	230	1,065	2,090
	400,000	595	2,700	5,280	230	1,065	2,090
	450,000	595	2,700	5,280	230	1,065	2,090

Conditions and procedures that occur as frequently as General Assessment will give at least 1000 people in five year age groups if sampling is done at a rate of 20% in Saskatchewan. For others that occur as frequently as Chickenpox, at least 100 people in the sample could

only be obtained if the total population of either males or females was considered. There are many conditions and procedures that do not occur as frequently as Chickenpox and so sampling at a rate of 20% will not guarantee having at least 100 people in the sample even if no age breakdown is considered.

Now look at the 50% sampling rate in Saskatchewan. Appendectomies require the following age groups to obtain at least 100 individuals in each group: under 15, 15-24, 25 and over. Any conditions or procedures that occur at the same frequency as Appendectomies would require very broad age groups to achieve even this small number of 100 individuals. Any conditions or procedures that are more infrequent would not allow any age breakdown to obtain at least 100 individuals in the sample that have the condition of interest, even for as large a sampling rate as 50%. There are quite a number of conditions or procedures fitting into this category and so sampling at 50% will not be sufficient to meet the specification that 100 people in the sample have the condition of interest.

It should be pointed out here that this specification is quite lenient. Consider the implications of what would happen if the relationship between two such low frequency conditions or procedures were to be studied. If 100 individuals in the sample have a particular condition of interest and the study is to look at the relationship between this condition and a procedure occuring at a rate of 10%, then about 10 people in the sample would be expected to have had both the condition and the procedure of interest. Any conclusion based on 10 people in a sample where the population size is upwards of 20,000 in each age and sex group is questionable.

Thus, method two has shown that sampling in Saskatchewan at the large rate of 50% will not provide enough individuals in age and sex groups for reliable studies of low frequency conditions and procedures.

Table XI shows the proportion P of the population N that is required to guarantee that at least m individuals in the sample have the condition of interest when the province of Ontario is sampled at a rate of 20%. Comparing Table XI to Table III, the proportion of the population of Saskatchewan having certain conditions and procedures, the effect of sampling in the larger provinces can be determined. Conditions and procedures that occur as frequently as Acute Nasopharyngitis will yield at least 500 individuals in the sample in each five year age group in Ontario. Those conditions and procedures that are relatively infrequent, such as Appendectomies, will yield at least 500 individuals in the sample in each age group of ten or fifteen years. In Saskatchewan, only 100 individuals in the sample would be guaranteed if ages were grouped in much larger categories. Thus, it is again shown that sampling at a rate of 20% in Ontario is feasible whereas sampling at any rate in Saskatchewan is not advisable.

TABLE XI: P(%) Required for Specified m when f = 0.2 and a = 1.64

a	N	f = 0.2		
		m=100	m=500	m=1000
		%	%	%
1.64	100,000	0.58	2.67	5.23
	200,000	0.29	1.33	2.64
	400,000	0.14	0.67	1.31
	600,000	0.10	0.44	0.87
	800,000	0.07	0.33	0.65
	1,000,000	0.06	0.27	0.52

Summary

Both of the two methods discussed have shown that sampling does not seem advisable for the provinces with populations the same size or smaller than that of Saskatchewan. Because of the fact that medical conditions and procedures are relatively infrequent, the estimated rates or numbers might not be reliable or precise enough. The other problem is that the number of individuals that have the characteristics under study in the sample might be so small that the study itself would be questionable, especially considering the migration of individuals in and out of provincial insurance schemes.

Therefore, sampling should not be considered for the six smaller population provinces at the initial start up of the data base. Perhaps in later years when some of the files are received and analyzed and a better idea of the types of studies that the data base will handle are known, the question of sampling the provinces can be looked at again.

Included in the discussions of the two methods was a comparison with sampling in Ontario. Although not discussed in great detail, it appears that the problem of estimating low frequency conditions and procedures will not deter from sampling in the larger population provinces.

Number/
Numéro

2

M064F
c.2

Statistics Canada / Statistique Canada
Ottawa

FIABILITE DE L'ECHANTILLONNAGE POUR LA BASE DE DONNEES SUR LES SOINS MEDICAUX RECUS PAR LES MALADES

K. McClean
Division des méthodes
d'enquête - institutions
et agriculture
décembre 1981

TABLE DES MATIÈRES

INTRODUCTION

RENSEIGNEMENTS GÉNÉRAUX

- Première méthode : coefficient de variation
- Deuxième méthode : nombre minimal de cas d'intérêt dans l'échantillon

RÉSUMÉ

TABLEAU I : Population assurée, Saskatchewan (30 juin 1973)

TABLEAU IIA : Nombre de malades, hommes, Saskatchewan 1973

TABLEAU IIB : Nombre de malades, femmes, Saskatchewan 1973

TABLEAU IIIA : Pourcentage de malades, hommes, Saskatchewan 1973

TABLEAU IIIB : Pourcentage de malades, femmes, Saskatchewan 1973

TABLEAU IV : Valeur approximative de NP requise pour obtenir un certain C.V. avec une valeur donnée de f et de N

TABLEAU V : Population au recensement de 1976 (arrondie au millier le plus proche), Ontario

TABLEAU VI : Valeur approximative de P (%) requise pour obtenir un certain C.V.

TABLEAU VII : Taille de l'échantillon tiré d'une population infinie, seuil de confiance de 95 %

TABLEAU VIII : Valeur approximative de la fraction de sondage (%) en fonction de la précision des estimations et de la taille de la population

TABLEAU IX : Taille (minimale) de la population (en milliers) pour diverses fractions de sondage et pour certains degrés de précision des estimations

TABLEAU X : Valeur approximative de NP requise pour certaines valeurs de m, f et a pour lesquelles $PR(r < m) < \varepsilon$, c'est-à-dire $Pr(\frac{r}{n} < \frac{m}{n}) < \varepsilon$ et

$$PR(x < a\,\sigma_{\frac{r}{n}}) = \varepsilon$$

TABLEAU XI : Valeur de P (%) requise pour certaines valeurs de m quand $f = 0.2$ et $a = 1.64$

INTRODUCTION

La base de données sur les soins médicaux reçus par les malades a été proposée dans le but d'étudier les soins dispensés aux particuliers, au cours de leur vie, pour différentes maladies. En moyenne, dix réclamations par année sont présentées à la régie de l'assurance-maladie par chaque bénéficiaire. Si le projet de base de données devait englober tous les assurés au Canada, il faudrait dépouiller quelque deux cent cinquante millions de réclamations chaque année. Afin de réduire ce nombre, il a été proposé que des échantillons de bénéficiaires soient prélevés dans les dix provinces.

La population des dix provinces s'échelonne entre 125,000 à l'Île-du-Prince-Édouard et 8,500,000 en Ontario. Cinq provinces comptent moins d'un million d'habitants. Le problème que posent ces provinces à faible population est que l'estimation des maladies et des actes médicaux peu fréquents risque de ne pas être fiable si elle est faite à partir d'échantillons.

Ce document examine l'application de deux méthodes pour l'estimation des maladies et des actes médicaux peu fréquents à partir d'échantillons tirés dans les provinces à faible population. La première méthode est basée sur le résultat pour coefficient de variation pour un echantillonnage donné tandis que la deuxième analyse la fréquence des maladies et des actes médicaux qui correspondrait à un nombre minimal de cas dans l'échantillon. Ces deux méthodes sont présentées séparément dans les sections suivantes.

RENSEIGNEMENTS GÉNÉRAUX

Au cours de la période de 1972 à 1975, Statistique Canada a mis sur pied un projet pilote en collaboration avec la Saskatchewan. Les dossiers des réclamations présentées à la Saskatchewan Medical Care Commission ont été transmis à Statistique Canada. Il a été facile d'obtenir des totalisations pour 1973 montrant le nombre de bénéficiaires ayant reçu au moins un service pour chaque affection codée à trois chiffres dans la CIMA-8 et au moins un service pour chaque acte médical figurant dans la table des honoraires. A cause de la disponibilité des données et du fait que la Saskatchewan est une des provinces à faible population, on a utilisé ces totalisations de 1973 pour le présent document.

Le tableau 1 montre la répartition des assurés en Saskatchewan en fonction de l'âge et du sexe, au 30 juin 1973. Tous les taux présentés pour la Saskatchewan dans ce document sont calculés à partir de cette estimation stable de la population. Le problème de l'évolution continuelle de la taille d'une population n'est pas abordé ici.

Tableau I : Population assurée, Saskatchewan (30 juin 1973)

	Age		
	Total	Moins de 1	75 et plus
Hommes			
Femmes			
TOTAL			

Certaines maladies ou certains actes médicaux qui pourraient faire l'objet d'études fondées sur la base de données ont été mentionnés au cours des réunions de l'équipe de ce projet. Les tableaux IIA et IIB indiquent respectivement le nombre d'hommes et de femmes qui ont reçu au moins un service relatif à ces maladies ou à ces actes, et les tableaux IIIA et IIIB présentent ces résultats en pourcentages, c'est-à-dire qu'on a divisé les nombres fournis au tableau II par les chiffres de population du tableau I.

Les affections et les actes qui figurent dans ces tableaux couvrent toute une gamme de fréquences et peuvent donc représenter toute affection ou tout acte médical qu'on pourrait vouloir étudier à partir de la base de données. Par exemple, l'examen complet par un médecin généraliste, subi par quinze à cinquante pour cent des personnes dans chaque groupe d'âge et sexe, peut représenter n'importe quelle affection ou acte médical très fréquents. À l'autre extrême, les psychoses alcooliques, la sclérose en plaques et les appendicectomies peuvent représenter des maladies et traitements dont la fréquence est moins de la moitié d'un pour cent. En outre, un certain nombre d'affections et de traitements qui figurent dans les tableaux peuvent représenter les cas observés soit uniformément dans tous les groupes d'âge soit plus particulièrement dans les groupes d'âge jeune, moyen ou avancé. Bref, même si seulement seize maladies et quatorze actes médicaux figurent dans les tableaux, ces éléments représentent une grande variété d'affections et de traitements.

- Première méthode : Coefficient de variation

Dans l'échantillonnage aléatoire simple, la proportion P de personnes qui ont une certaine maladie dans la population ou dans un sous-groupe d'âge et sexe de la population de taille N est estimée par la proportion p de ce genre de malades dans un échantillon de taille n. L'estimateur du nombre de personnes dans la population qui ont cette maladie est donc $\hat{A} = Np$ et $\mathrm{Var}(\hat{A}) = \frac{N^2 P(1-P)}{n} \left(\frac{N-n}{N-1}\right)$. Étant donné qu'il s'agit ici de populations assez grandes, on peut utiliser l'approximation $N-1 \simeq N$ pour simplifier la formule de la variance :

$$\text{Var}(\hat{A}) =$$

où $f = \frac{n}{N}$ est la fraction de sondage. Le coefficient de variation de $\hat{A}$, qui est désigné par C.V., est le résultat de l'équation

$$\text{C.V.} =$$

(1)

TABLEAU IIA : Nombre de malades, hommes, Saskatchewan 1973

MALADIES ET ACTES MEDICAUX			AGE
	Code	Description	
CIMA - 8e RÉVISION	250	Diabète sucré	
	400-404	Maladies hypertensives	
	410-414	Maladie ischémique du coeur	
	052	Varicelle	
	056	Rubéole	
	075	Mononucléose infectieuse	
	291	Psychose alcoolique	
	340	Sclérose en plaques	
	460	Rhino-pharyngite aiguë (rhume)	
	463	Amygdalite aiguë	
	540-543	Appendicite	
	571	Cirrhose du foie	
	640-645	Avortement	
	640	Avortement thérapeutique	
	643	Avortement spontané	
	749	Fissure du palais et bec-de-lièvre	
TABLE DES HONORAIRES	3B	Examen complet - MG	
	87L	Mastectomie radicale	
	241L	Réparation d'une hernie hiatale	
	333L	Appendicectomie	
	499L	Excision ou cautérisation d'une verrue plantaire unique	
	359M	Arthroscopie	
	320N-329N	Réparation, fissure labiale & fente palatine	
	35P	Amniocentèse trans-abdominale	
	118P	Dilatation & curetage	
	120P-125P	Hystérectomie	
	65R-66R	Circoncision, nouveaux-nés & jeunes enfants	
	190R	Vasectomie	
	145T-146T	Amygdalectomie	
	131A	Spécimen pour cytologie vaginale	

TABLEAU IIB : Nombre de malades, femmes, Saskatchewan 1973

MALADIES ET ACTES MÉDICAUX		ÂGE
	Code	Description
CIMA - 8e RÉVISION	250	Diabète sucré .
	400-404	Maladies hypertensives
	410-414	Maladie ischémique du coeur
	052	Varicelle
	056	Rubéole
	075	Mononucléose infectieuse
	291	Psychose alcoolique
	340	Sclérose en plaques
	460	Rhino-pharyngite aiguë (rhume)
	463	Amygdalite aiguë
	540-543	Appendicite
	571	Cirrhose du foie
	640-645	Avortement
	640	Avortement thérapeutique
	643	Avortement spontané
	749	Fissure du palais et bec-de-lièvre
TABLE DES HONORAIRES	3B	Examen complet - MG
	87L	Mastectomie radicale
	241L	Réparation d'une hernie hiatale
	333L	Appendicectomie
	499L	Excision ou cautérisation d'une verrue plantaire unique
	359M	Arthroscopie
	320N-329N	Réparation, fissure labiale & fente palatine
	35P	Amniocentèse trans-abdominale
	118P	Dilatation & curetage
	120P-125P	Hystérectomie
	65R-66R	Circoncision, nouveaux-nés & jeunes enfants
	190R	Vasectomie
	145T-146T	Amygdalectomie
	131A	Spécimen pour cytologie vaginale

TABLEAU IIIA : Pourcentage de malades, hommes, Saskatchewan 1973

	MALADIES ET ACTES MÉDICAUX		ÂGE
	Code	Description	
CIMA - 8e RÉVISION	250	Diabète sucré	
	400-404	Maladies hypertensives	
	410-414	Maladie ischémique du coeur	
	052	Varicelle	
	056	Rubéole	
	075	Mononucléose infectieuse	
	291	Psychose alcoolique	
	340	Sclérose en plaques	
	460	Rhino-pharyngite aiguë (rhume)	
	463	Amygdalite aiguë	
	540-543	Appendicite	
	571	Cirrhose du foie	
	640-645	Avortement	
	640	Avortement thérapeutique	
	643	Avortement spontané	
	749	Fissure du palais et bec-de-lièvre	
TABLE DES HONORAIRES	3B	Examen complet - MG	
	87L	Mastectomie radicale	
	241L	Réparation d'une hernie hiatale	
	333L	Appendicectomie	
	499L	Excision ou cautérisation d'une verrue plantaire unique	
	359M	Arthroscopie	
	320N-329N	Réparation, fissure labiale & fente palatine	
	35P	Amniocentèse trans-abdominale	
	118P	Dilatation & curetage	
	120P-125P	Hystérectomie	
	65R-66R	Circoncision, nouveaux-nés & jeunes enfants	
	190R	Vasectomie	
	145T-146T	Amygdalectomie	
	131A	Spécimen pour cytologie vaginale	

TABLEAU IIIB : Pourcentage de malades, femmes, Saskatchewan 1973

MALADIES ET ACTES MÉDICAUX			ÂGE
	Code	Description	
CIMA - 8e RÉVISION	250	Diabète sucré	
	400-404	Maladies hypertensives	
	410-414	Maladie ischémique du coeur	
	052	Varicelle	
	056	Rubéole	
	075	Mononucléose infectieuse	
	291	Psychose alcoolique	
	340	Sclérose en plaques	
	460	Rhino-pharyngite aiguë (rhume)	
	463	Amygdalite aiguë	
	540-543	Appendicite	
	571	Cirrhose du foie	
	640-645	Avortement	
	640	Avortement thérapeutique	
	643	Avortement spontané	
	749	Fissure du palais et bec-de-lièvre	
TABLE DES HONORAIRES	3B	Examen complet - MG	
	87L	Mastectomie radicale	
	241L	Réparation d'une hernie hiatale	
	333L	Appendicectomie	
	499L	Excision ou cautérisation d'une verrue plantaire unique	
	359M	Arthroscopie	
	320N-329N	Réparation, fissure labiale & fente palatine	
	35P	Amniocentèse trans-abdominale	
	118P	Dilatation & curetage	
	120P-125P	Hystérectomie	
	65R-66R	Circoncision, nouveaux-nés & jeunes enfants	
	190R	Vasectomie	
	145T-146T	Amygdalectomie	
	131A	Spécimen pour cytologie vaginale	

Si l'on calcule P à l'aide de l'équation (1), on obtient

$$P = \qquad (2)$$

À l'aide de l'équation (2), on peut calculer, pour la population de taille N, la proportion P de personnes qui ont une maladie donnée lorsque la fraction d'échantillonnage est f. Le tableau IV indique le nombre de malades NP nécessaire pour obtenir un coefficient de variation donné à partir d'un échantillon prélevé d'une population de taille N selon une fraction de sondage f. Par exemple, dans le cas d'un échantillon tiré d'une population où N = 100,000 et où la fraction de sondage est de 0.10, il faudrait, pour que le coefficient de variation soit égal ou inférieur à 0.1, qu'au moins 890 de ces 100,000 personnes aient la maladie en cause.

TABLEAU IV : Valeur approximative de NP requise pour obtenir un certain C.V. avec une valeur donnée de f et de N

Le coefficient de variation est une mesure de la fiabilité de l'estimateur $\hat{A}$. Les estimateurs pour lesquelles C.V. < 0.05, c'est-à-dire celles dont l'erreur représente moins de 5 % de la vraie valeur, sont assez fiables, tandis que les estimateurs pour lesquelles C.V. > 0.2, celles dont l'erreur n'est pas moins de 20 % de la valeur réelle, sont généralement peu fiables. En comparant le tableau IV aux tableaux IIA et IIB, on peut voir pour quelles maladies ou actes médicaux on pourrait obtenir des estimateurs fiables si on avait prélevé un échantillon en Saskatchewan.

Dans le cas de maladies ou d'actes médicaux aussi fréquents que l'examen complet, le coefficient de variation serait de moins de 0.05, même lorsque la fraction de sondage est de seulement 5 %. Par contre, s'il s'agit de maladies ou d'actes aussi peu fréquents que la sclérose en plaques, le coefficient de variation dépasserait 0.2, même pour une fraction de sondage de 50 %.

Pour un bon nombre des maladies et des actes médicaux qui figurent dans les tableaux IIA et IIB, une fraction de sondage de 5 % n'entraînerait pas un coefficient de variation de moins de 0.10, même si l'on voulait étudier seulement la population masculine ou féminine sans recoupement en fonction de l'âge. Étant donné que l'âge est un facteur qui présente beaucoup d'intérêt dans les études médicales, une fraction de sondage si faible n'est pas à conseiller.

Une fraction de sondage de 10% ou de 20% est meilleur pour la Saskatchewan qu'une fraction de 5 %, mais quelques maladies et actes médicaux sont tellement rares qu'il est impossible d'atteindre un coefficient de variation de moins de 0.2, même pour l'ensemble de la population masculine ou féminine. Dans un certain nombre de catégories, on obtient un coefficient de variation de moins de 0.10 avec une fraction de sondage de 20 % en formant des groupes d'âge de dix ou vingt ans mais, pour ces mêmes groupes, c'est seulement dans le cas des maladies et des traitements assez fréquents que C.V. < 0.10 lorsque la fraction de sondage est de 10 %.

Une fraction de sondage de 50 % en Saskatchewan produira un C.V. < 20 pour l'ensemble de la population masculine ou féminine, dans le cas des maladies et des actes médicaux aussi rares que la réparation d'une fissure du palais ou d'un bec-de-lièvre. Quant aux autres affections et traitements, il semble suffisant de former des groupes d'âge de cinq ou de dix ans pour que C.V. < 0.10.

Cette méthode démontre que des échantillons prélevés dans les six provinces les moins populeuses, selon une fraction de sondage de 50 %, ne fourniront pas des estimateurs fiables pour des petits groupes d'âge et pour toute la gamme de maladies et d'actes médicaux qu'on voudrait éventuellement étudier. L'estimation ne devrait donc pas se faire à partir d'échantillons dans ces provinces.

Examinons le cas de l'Ontario, la province la plus populeuse. La répartition de la population de cette province est présentée au tableau V. Au moyen de l'équation (2), on obtient le tableau VI, qui montre la proportion approximative de la population requise pour obtenir certaines valeurs du coefficient de variation et pour quelques valeurs données de la taille de la population et de la fraction de sondage. Si l'on suppose que la fréquence relative des maladies et des actes médicaux en Ontario est la même qu'en Saskatchewan, une comparaison du tableau VI et du tableau III révèle que l'échantillonnage produit des estimateurs plus fiables dans le cas de l'Ontario. Avec une fraction de sondage de 20 % en Ontario, les estimateurs relatives à des maladies et à des actes médicaux aussi rares que la psychose alcoolique, la sclérose en plaques et l'appendicectomie auront un coefficient de variation de moins de 0.2 pour des groupes de cinq années d'âge et, dans certains cas, ce coefficient sera inférieur à 0.1. Il ressort donc que, même si l'échantillonnage ne convient pas aux provinces dont la population est moins grande que celle de la Saskatchewan, il peut être fait dans les provinces les plus populeuses.

TABLEAU V : Population au recensement de 1976 (arrondie au millier le plus proche), Ontario

	TOTAL	AGE Moins de 1	75 et plus
Hommes			
Femmes			
Total			

TABLEAU VI : Valeur approximative de P (%) requise pour obtenir un certain C.V.

Il existe une autre façon de vérifier les résultats de la première méthode, qui est décrite dans le document "Adequacy of Sample Size" de la Sous-section de la méthodologie de la statistique de la santé de l'Organisation mondiale de la santé à Genève. Dans cette autre technique, il faut calculer la fraction de sondage requise pour un échantillon tiré d'une population infinie et appliquer un facteur correctif à cette fraction afin de trouver la fraction de sondage nécessaire pour constituer un échantillon de la population finie qu'on veut étudier. Les trois éléments qui définissent la taille requise de l'échantillon sont :

1. la précision recherchée de l'estimation (c'est-à-dire le pourcentage d'erreur permis)

2. le seuil de confiance voulu (par exemple, une probabilité de 95 % que la précision recherchée est atteinte)

et 3. une hypothèse a priori de la proportion de la population qui possède la caractéristique d'intérêt.

Le tableau VII indique l'intervalle de valeurs permises autour du vrai pourcentage dans quinze cas, ainsi que la taille requise d'un échantillon tiré d'une population infinie afin que la vraie valeur soit dans l'intervalle donné avec un seuil de confiance de 95 %. En appliquant le facteur correctif aux chiffres du tableau VII, on obtient les fractions de sondage approximatives nécessaires pour atteindre la précision recherchée autour d'un estimé calculé pour des populations finies de diverses tailles, et les résultats sont présentés au tableau VIII. Par exemple, pour avoir une confiance a 95 % que la vraie valeur se trouve dans un intervalle de 5 % autour d'une proportion estimée de 10 % pour une population de 40,000 personnes, la fraction de sondage requise serait de 35 % et la taille de l'échantillon serait 0.35 x 40,000 = 14,000.

Afin de déterminer la taille la plus petite d'une population finie, qui est suffisante pour assurer la précision recherchée autour des estimateurs qui correspondent à différentes fractions de sondage, on a remanié le tableau VIII sous la forme du tableau IX. Ainsi, pour obtenir une confiance de 95 % que la vraie valeur se trouve dans un intervalle de 10 % autour de l'estimé 0.05, la taille de la population doit être d'au moins 25,000 si la fraction de sondage est de 25 %.

Le document de l'OMS n'incluait pas des échantillons de taille supérieure à 100,000 personnes; il a donc été impossible d'appliquer un facteur correctif à ce genre de

TABLEAU VII : Taille de l'échantillon tiré d'une population infinie, seuil de confiance de 95 %

	1 % de précision		5 % de précision		10 % de précision	
Estimations	Intervalle	Taille de la pop. infinie	Intervalle	Taille de la pop. infinie	Intervalle	Taille de la pop. infinie

cas afin de calculer la fraction d'échantillonnage de la population finie. C'est pour cette raison qu'on retrouve >50 % dans certaines cases du tableau VIII.

Par conséquent, il a été impossible de calculer la plus petite taille de la population pour l'estimé 0.001, quelle que soit la fraction de sondage, ainsi que pour les l'estimé inférieures a 0.01 avec 5 % de précision et pour les estimés de moins de 0125 avec 1 % de précision. On peut seulement dire que ces tailles de population seraient plus élevées que celles se rapportant aux grandes estimés.

Pour 10 % de précision, il semble que l'échantillonnage serait possible dans le cas d'estimé de 0.005 ou plus. Si la fraction de sondage est de 15 % ou plus, on pourrait calculer des estimés de 0.07 ou plus avec 10 % de précision et un seuil de confiance de 95 % à partir de populations d'au moins 40,000 personnes, notamment pour les groupes de cinq années d'âge en Saskatchewan. Il serait même possible d'obtenir des estimés aussi petites que 0.005 dans le cas de populations de 100,000 personnes avec une fraction de sondage de 50 %. En ce qui concerne la Saskatchewan, cela reviendrait à former des groupes d'âge de vingt à vingt-cinq ans. Toutefois, le tableau III indique qu'un certain nombre de maladies et d'actes médicaux ont une fréquence inférieure à 0.5 %, et ces estimés.

n'atteindraient donc pas une précision de 10 % si la fraction de sondage était de 50 % et si les populations comptaient 100,000 personnes.

Cette deuxième technique mène aux mêmes conclusions que celles de la première méthode, à savoir qu'il serait possible de prélever des échantillons dans des provinces de même taille que la Saskatchewan à condition de faire des estimations seulement pour les maladies et les actes médicaux de fréquence relativement élevée. Les catégories aussi peu fréquentes que la mononucléose infectieuse, la sclérose en plaques ou les appendicectomies ne pourraient pas être étudiées, même avec une fraction de sondage de 50 %.

TABLEAU VIII : Valeur approximative de la fraction de sondage (%) en fonction de la précision des estimations et de la taille de la population

Estimation	Précision

TABLEAU IX : Taille (minimale) de la population (en milliers) pour diverses fractions de sondage et pour certains degrés de précision des estimations

Précision	Fraction de sondage	ESTIMATIONS

Degré de précision

o Deuxième méthode : Nombre minimal de cas d'intérêt dans l'échantillon

Cette méthode permet de déterminer P, la proportion de personnes dans la population de taille N qui possèdent une caractéristique d'intérêt, sous la contrainte que, pour la fraction de sondage f, la probabilité d'observer moins de m personnes ayant cette caractéristique dans l'échantillon de taille n soit très petite. Soient R et r respectivement le nombre de personnes qui possèdent la caractéristique d'intérêt dans la population et dans l'échantillon. Quand n est assez grand, on peut démontrer que $\frac{r}{n} \sim$ Normal $(P, \frac{P(1-P)}{n} \frac{(N-n)}{N-1})$. Étant donné qu'il s'agit ici de grandes populations, l'approximation $N-1 \simeq N$ permet d'obtenir la distribution $\frac{r}{n} \sim$ Normal $(P, \frac{P(1-P)(1-f)}{n})$. Pour satisfaire à la condition $Pr(r < m) < \varepsilon$ ou $Pr(\frac{r}{n} < \frac{m}{n}) < \varepsilon$, il faut trouver la valeur de P dans l'équation suivante :

$$\frac{m}{n} = P - a \sqrt{\frac{P(1-P)(1-f)}{n}} \qquad (3)$$

où $Pr(X < a \sigma_{\frac{r}{n}}) \leq \varepsilon$ et $X \sim$ Normal (0,1). Si on calcule P à l'aide de l'équation (3), on obtient

$$P = \qquad (4)$$

L'équation (4) fournit deux valeurs de P, P_1 et P_2, de sorte que $P_2 > P_1$. Examinons la distribution de $\frac{r}{n}$.

La zone hachurée représente la probabilité décrite plus haut, c'est-à-dire $Pr(\frac{r}{n} < \frac{m}{n}) = \mathcal{E}$. Il s'ensuit donc que P_2 est la valeur d'intérêt et qu'il faut seulement utiliser le signe plus dans la formule (4).

Prenons les deux valeurs suivantes pour $\mathcal{E}$: 0.05 et 0.025. Il en découle deux valeurs de a, qu'on obtient ainsi :

A partir de ces deux valeurs de a et avec diverses valeurs de m, f et N, on peut calculer P pour différentes combinaisons des variables de l'équation (4). Le tableau X présente les valeurs de NP, le nombre de personnes dans la population qui possèdent la caractéristique d'intérêt, de sorte que NP satisfait à la contrainte selon laquelle au moins m personnes dans l'échantillon auront cette caractéristique avec une certaine probabilité.

TABLEAU X : Valeur approximative de NP requise pour certaines valeurs de m, f et a pour lesquelles $Pr(r < m)$, c'est-à-dire $Pr(\frac{r}{n} < \frac{m}{n}) < \mathcal{E}$ et $Pr(x < a\,\sigma_{\frac{r}{n}}) = \mathcal{E}$.

Les affections et les actes médicaux aussi fréquents que l'examen complet concernera au moins mille personnes dans chaque groupe de cinq années d'âge, si la fraction de sondage est de 20 % en Saskatchewan. Quant aux maladies aussi peu fréquentes que la varicelle, un échantillon comporterait au moins 100 personnes seulement qui ont cette caractéristique si on n'étudiait que l'ensemble de la population masculine ou féminine. Un certain nombre de maladies et d'actes médicaux sont encore plus rares que la varicelle et une fraction de sondage de 20 % ne garantit par l'inclusion d'au moins 100 personnes concernées dans l'échantillon, même s'il n'est pas question de faire un recoupement en fonction de l'âge.

Examinons maintenant le cas d'une fraction de sondage de 50 % en Saskatchewan. Pour les appendicectomies, il faut sélectionner au moins 100 personnes dans les groupes d'âge suivants : moins de 15 ans, de 15 à 24 ans, 25 ans et plus. Chaque maladie ou acte médical aussi fréquent que les appendicectomies exigerait donc des groupes d'âge très étendus afin qu'on puisse sélectionner même juste 100 personnes. Dans le cas des maladies ou actes médicaux qui sont encore plus rares, il serait impossible de sélectionner au moins 100 personnes concernées dans n'importe quel groupe d'âge, même si la fraction de sondage était de 50 %. Un bon nombre de maladies et d'actes médicaux sont assez peu fréquents et, par conséquent, un taux de sondage de 50 % ne permettrait pas de satisfaire à la contrainte d'inclure au moins 100 personnes concernées dans l'échantillon.

Il faut souligner que cette contrainte n'est pas très exigente. Imaginons ce qui arriverait si on voulait étudier la relation entre deux maladies ou actes médicaux qui sont peu fréquents. Si 100 personnes dans un échantillon ont une affection particulière et si l'on veut examiner le lien entre cette maladie et un acte médical qui est posé dans 10 % des cas, on s'attendrait à observer qu'une dizaine de personnes dans l'échantillon ont eu la maladie en question et subi le traitement correspondant. Toute conclusion fondée sur l'observation de 10 personnes dans un échantillon qui représente une population de plus de 20,000 personnes dans chaque groupe d'âge et sexe est douteuse.

Ainsi, la deuxième méthode montre que, même avec une fraction de sondage assez élevée, 50 %, l'échantillonnage en Saskatchewan ne permettra pas de sélectionner un nombre assez grand de personnes dans chaque groupe d'âge et sexe pour faire des études fiables des maladies et des actes médicaux peu fréquents.

Le tableau XI indique la proportion P de la population N, qui est nécessaire pour qu'au moins m personnes dans l'échantillon aient la caractéristique d'intérêt lorsqu'on applique une fraction de sondage de 20 % en Ontario. En comparant ce tableau au tableau III, qui montre la proportion de la population de la Saskatchewan qui a eu certaines maladies ou subi certains traitements, on peut décrire l'effet de l'échantillonnage dans les provinces populeuses. Dans le cas de maladies et d'actes médicaux aussi fréquents que la rhino-pharyngite aiguë, il serait possible de sélectionner au moins 500 personnes concernées dans l'échantillon de chaque groupe de cinq années d'âge, en Ontario. Quant aux maladies et aux actes médicaux relativement peu fréquents, comme les appendicectomies, on obtiendrait au moins 500 personnes concernées dans chaque groupe de dix ou de quinze années d'âge. En Saskatchewan, on obtiendrait seulement 100 personnes concernées dans l'échantillon de chaque groupe d'âge, à condition que ces catégories soient beaucoup plus étendues. Donc ici encore, on conclut qu'il serait possible d'utiliser une fraction de sondage de 20 % en Ontario, tandis qu'aucune fraction ne convient dans le cas de la Saskatchewan.

TABLEAU XI : Valeur de P (%) requise pour certaines valeurs de m quand f = 0.2 et a = 1.64

Résumé

Les deux méthodes décrites dans les sections précédentes indiquent qu'il ne serait pas approprié de prélever des échantillons dans les provinces où la taille de la population est égale ou inférieure à celle de la Saskatchewan. Dans le cas de maladies et d'actes médicaux relativement peu fréquents, les proportions ou les nombres estimés risqueraient de ne pas être fiables ou assez précis. L'autre problème est que le nombre de personnes qui possèdent une caractéristique d'intérêt pourrait être si faible que les résultats seraient douteux, surtout si l'on pense au phénomène des assurés qui changent de régime provincial d'assurance-maladie par la migration.

Il ne faudrait donc pas prélever des échantillons dans les six provinces les moins populeuses, à l'étape de la construction de la base de données. Il est possible que, dans quelques années, lorsqu'on aura reçu et analysé certains fichiers et qu'on aura une meilleure idée du genre d'étude qui sera faite à partir de la base de données, la question de l'échantillonnage sera examinée de nouveau.

Dans la description des deux méthodes présentées plus haut, on a comparé les résultats de l'échantillonnage en Saskatchewan et en Ontario. Bien que le problème de l'estimation des maladies ou des actes médicaux peu fréquents n'ait pas été analysé en détail ici, il semble que rien n'empêche l'échantillonnage dans les provinces les plus populeuses.